Reliability of the Craniomandibular Index

John P. Hatch, PhD
Professor
Departments of Psychiatry and Orthodontics

John D. Rugh, PhD
Professor and Chair
Department of Orthodontics

Shiro Sakai, DDS, MS
Assistant Professor
Department of Orthodontics

Thomas J. Prihoda, PhD
Associate Professor
Departments of Pathology and Psychiatry

The University of Texas Health Science Center at San Antonio
San Antonio, Texas

Correspondence to:
Dr John P. Hatch
Department of Orthodontics
The University of Texas Health Science Center at San Antonio
Mail Code 7910
7703 Floyd Curl Drive
San Antonio, TX 78229-3900
Fax: +(210) 567-6941
E-mail: hatch@uthscsa.edu

Aims: *To examine various dimensions of reliability of the Craniomandibular Index, a commonly used instrument for quantifying the severity of signs and symptoms of temporomandibular disorders.* ***Methods:*** *Classical psychometric theory and generalizability theory were used to assess the reliability of data obtained from a calibration study of examiners participating in a multi-site clinical trial and from a random community sample.* ***Results:*** *The reliability of aggregate scores formed by summing individual binary scored items was high, with intraclass correlations ranging from 0.81 to 0.88. When it was required that examiners recognize and agree upon a specific pattern of signs and symptoms exhibited by a patient, however, reliability dropped dramatically (multivariate kappas ranged from 0.26 to 0.32). A group of practicing examiners also showed limited ability to agree with the pattern of signs and symptoms identified by a "gold standard" examiner (multivariate kappas ranging from 0.25 to 0.32). Generalizability analysis failed to identify the specific sources of measurement error that played a major role in limiting reliability but demonstrated that generalizability of aggregate scores was very high.* ***Conclusion:*** *Methods of classical psychometric theory and generalizability theory support the conclusion that the reliability of aggregate scores is acceptably high. Individual items assessing certain aspects of jaw mobility and joint sounds are measured with poor reliability. Reliability declines when it is defined as the ability of examiners to agree among themselves upon a specific constellation of signs and symptoms or their ability to identify correctly a "correct" constellation identified by an expert examiner.*
J OROFAC PAIN 2002;16:284–295.

Key words: Craniomandibular Index, temporomandibular disorders, reproducibility of results, psychometrics, generalizability theory

Advancement of our understanding of temporomandibular disorders (TMD) is limited by our current ability to specify and apply diagnostic criteria reliably for these disorders and their clinical subtypes. Discovery of causative factors and evaluation of treatment techniques depend on clinicians' ability to identify and classify individual signs and symptoms. In both clinical and research settings it is important that the severity with which signs and symptoms are manifested be reliably measurable. Currently, the defining symptoms of TMD include pain, limitation of mandibular movement, and/or sounds emanating from the temporomandibular joint (TMJ). Various techniques have been advanced to assess these symptoms.[1–3] Each of these techniques involves making similar clinical measurements. To be used with confidence, measurement techniques must produce equivalent results when used by different examiners or on different occasions or in different settings. In other words, they must demonstrate good reliability.

Several efforts have used classical psychometric methods[4–8] to assess the reliability of various TMD examination formats. These studies show, in general, that measurement of jaw movement with a millimeter ruler can be performed with acceptable reliability, but measurement of muscle and joint palpation pain and joint sounds often can be measured with only modest to poor reliability.

Classical psychometric theory asserts that an observed score is composed of 2 components of variability. One component is assumed to be systematic and is referred to as *true score variance.* In assessing TMD, true score variance would be the variation among patients because they truly possess to varying degrees the clinical attributes being assessed. The second variance component is assumed to be random and represents measurement error. When true score variance is high relative to measurement error, we say that a measurement is reliable. This is an elegantly simple and useful conceptualization. However, since measurement error is not a monolithic construct, several definitions of reliability are traditionally used under classical theory. If measurement error due to disagreement among examiners is the concern, then we speak of *interexaminer reliability.* If measurement error due to measurement occasion is the concern, then we speak of *test-retest reliability.* If measurement error in item sampling is the concern, then we speak of *internal consistency.* Classical psychometric theory requires a piecemeal assessment of these different breeds of reliability and renders reliability a rather mercurial concept.

Generalizability theory is a method that seeks to partition measurement error into multiple components and assess their relative impact on the measurement process.[9–11] Generalizability theory thus extends classical reliability theory in a very useful way. It allows the researcher to examine multiple sources of measurement error simultaneously. The concept of reliability is replaced by the more flexible concept of *generalizability,* which is the extent to which a measurement can be generalized to a wider set of conditions and circumstances called the *universe of generalization.*

The purpose of the present study was to apply both classical psychometric methods and modern generalizability theory to a TMD assessment scale, the Craniomandibular Index (CMI).[1,12] We report the results of 3 reliability studies using data from a large community-based epidemiologic study and an examiner calibration study conducted within a multi-site clinical trial. The interrater reliability and test-retest reliability of the CMI have been previously studied by classical methods and found to be high, with intraclass correlation coefficients (ICCs) ranging from 0.84 to 0.96 for the aggregate scores.[12]

Materials and Methods

Study 1: Internal Consistency

In the first study, interexaminer variability was held constant by having all subjects examined by a single calibrated and highly trained examiner. Since the subjects of this study were a random sample of community-dwelling middle-aged and elderly adults, signs and symptoms of TMD occurred at relatively low rates. The reliability estimates resulting from this study would be most applicable to epidemiologic studies involving nonclinical community samples.

Examiner. All subjects were examined by 1 well-trained and highly experienced dentist (SS) who had completed postgraduate training in orofacial pain management and is a diplomate of the American Board of Orofacial Pain. During this study he regularly performed the CMI examination as the primary clinical examiner for 2 large research studies and as the primary attending dentist in an active, university-based facial pain clinic. During the time that these data were being collected he received annual retraining and calibration from the developer of the CMI (Dr James Fricton; additional details provided below).

CMI Examination. The CMI was developed to quantify objectively the severity of signs and symptoms of TMD in the context of epidemiologic and clinical outcome studies.[1,12] It currently is widely used for that purpose. For this study, the examination methods were as described by Fricton and Schiffman and as taught by Fricton during annual training sessions.[12] The CMI consists of 62 items that are coded as positive or negative. In addition to an overall aggregate score (CMI), the CMI produces a joint Dysfunction Index (DI) and a Muscle Index (MI). The DI quantifies mandibular range of motion; TMJ noises such as clicking, popping, and crepitus; and tenderness of TMJ structures to manual palpation. The MI quantifies tenderness of facial, neck, and shoulder muscles to manual palpation. Patients were seated in a dental chair during examinations. While individual signs and symptoms were quantified, no attempt was made to establish a TMD diagnosis.

Subjects. Subjects were 913 individuals (403 men and 510 women) between the ages of 37 and 82 years (mean = 61.1, SD = 11.1) who were participants in the Oral Health: San Antonio Longitudinal Study of Aging. Data were collected between July 1994 and March 1998. Subjects were selected by a stratified random selection procedure from 3 socioculturally diverse neighborhoods of San Antonio, Texas: *(1)* a low-income, nearly exclusively Mexican-American "barrio" neighborhood; *(2)* a middle-income, "transitional" mixed Mexican-American/European-American neighborhood; and *(3)* an upper-income "suburban" neighborhood containing approximately 10% Mexican-Americans and 90% European-Americans. Subjects were excluded from the study only if they were pregnant or if their ethnic group could not be classified as either Mexican-American or European-American.

Data Analysis. The internal consistency of the CMI was estimated with Cronbach's alpha statistic. This statistic estimates the lower boundary of reliability for a test that is scored by summing the item scores, and can be interpreted as the correlation between the CMI items and all other scales constructed to measure the same phenomenon that contain the same number of items. Alpha can range from a negative value to unity. Cronbach's alpha depends to some degree on the number of items on the test under consideration and should be interpreted relative to other tests with a similar number of items. Pearson correlation coefficients were also calculated between the individual CMI items and the aggregate DI, MI, and CMI scores. These coefficients permit assessment of how individual items relate to the CMI aggregate scores.

Study 2: Interexaminer Agreement

In the second study, data were collected during annual sessions designed to retrain and calibrate clinical examiners participating in a long-term (9-year) multi-site clinical trial. For the calibration study, both patients currently receiving treatment for TMD and healthy volunteers were examined. The reliability estimates derived from this study would be most applicable to the clinical research setting.

Examiners. The subjects were 13 male and 9 female clinical examiners who were participants in the multi-site clinical trial. Fifteen examiners were dentists and 7 were dental hygienists. One examiner, Dr James Fricton, trained and re-calibrated the other examiners and served as the "gold standard" examiner. Examiners were affiliated with 4 universities: The University of Texas Health Science Center at San Antonio (n = 10), Emory University (n = 6), the University of Florida (n = 5), and the University of Minnesota (n = 1). Training/calibration sessions were conducted annually at San Antonio between 1990 and 1998, with the exception of 1994 (8 sessions total). Each study site designated 1 primary examiner and 1 or more backup examiners. Primary examiners performed CMI examinations routinely and on a regular basis, while backup examiners filled in when primary examiners were unavailable.

Patients and Controls. The individuals examined consisted of 98 women and 8 men who either were currently undergoing treatment for TMD at the Facial Pain Clinic of the University of Texas Health Science Center at San Antonio (n = 55) or were volunteers not currently undergoing treatment for TMD (n = 51). Patients and controls were paid for serving in the calibration study. Information about treatment history was not recorded.

Training. The training program has been described elsewhere in detail.[13] Briefly, training consisted of the examiner first watching a 19-minute instructional videotape then witnessing a demonstration by the training examiner of the proper examination technique. Each examiner performed a guided examination of the trainer and received instructional feedback. The training examiner and trainees then each examined 4 to 7 patients, and their results were compared and discussed. In each year, some examiners received training for the first time, while others received annual retraining. Because examiners periodically left the study and were replaced, all examiners did not examine all patients. The time required for training was approximately 5 hours.

Calibration. Following the training phase, all examiners, including the training examiner, examined approximately 12 patients independently and in random order and recorded their findings on standard forms. Privacy was maintained by placing each patient in a dental operatory and having examiners rotate among the patients who were seated in dental chairs. Examiners were blind to the other examiners' scores.

Data Analysis. The data submitted to analysis comprised a total of 574 CMI examinations. Mean DI, MI, and CMI scores were compared, across various subgroups of patients and examiners, by the Student *t* test and analysis of variance (ANOVA). Interrater agreement on individual binary coded items (ie, positive or negative) taken over the group of examiners was assessed by a generalized kappa statistic, a chance-corrected percent agreement measure.[14] Generalized agreement taken over all items contributing to aggregate scores was assessed by the method of Berry and Mielke.[15,16] Agreement between the group of study examiners and the "gold standard" examiner was measured by the use of a multivariate, chance-corrected measure designed for this purpose.[17] Interrater reliability of quasi-continuous, aggregate scores constructed by summing individual items (eg, MI, DI, and CMI) was assessed by ICCs.[18] The ICC is interpreted as the proportion of total variance attributable to true differences among the individuals examined, as opposed to other sources, including examiners. ICCs can range from negative values to 1.0, but when negative values occur they are interpreted to indicate zero interexaminer agreement. Because the design was unbalanced (ie, not all examiners examined all patients), variance components were estimated from a restricted maximum likelihood procedure rather than ANOVA. The ICCs were estimated from a random effects model. In other words, we estimated the reliability of a score obtained on a single patient randomly selected from the population of all such patients examined by a single examiner randomly selected from the population of examiners. This method also assumes that the scores of multiple examiners are not averaged to raise reliability.

Study 3: Generalizability

Subjects, Examiners, and Examination Methods. The data used in Study 2 were used again in this study.

Study Design. The objects of measurement in this study were the 106 subjects who underwent the CMI examination. The variance component associated with the objects of measurement corresponds to true score variability in classical psychometric theory. In addition, several additional dimensions of variability, called *facets* in generalizability theory, were examined. Each of these facets corresponds to a component of variability and represents a source of measurement error that may limit reliability. The first facet examined here represents individual differences among the 106 subjects who participated in the study. This facet represents variability resulting from random individual differences among the subjects examined and is analogous to true score variance under classical psychometric theory. In addition, however, variability was further partitioned.

It was hypothesized that the regularity with which an examiner examines patients might affect his ability to perform the CMI with high reliability. Therefore, primary versus backup status was included as a second facet of interest. Primary examiners performed the CMI on a regular basis at their study site, while backup examiners filled in when primary examiners were unavailable. The third facet represents potential measurement error due to the examiner's sex. During calibration sessions, casual observation suggested that differences between male and female examiners in fingertip anatomy and applied palpation pressure might affect digital palpation results. The fourth facet examined represents study site. Although all the individuals examined in this study were drawn only from the San Antonio site, the examiners represented 4 different university clinics. We hypothesized that examiners at different sites might adopt locally idiosyncratic examination techniques that could limit interexaminer agreement. The contributions of this facet would have implications for the generalizability of results from multi-site studies. The fifth facet corresponds to the examiner's professional training, ie, whether they were a dentist or a dental hygienist.

Data Analysis. Variance components were calculated by the VARCOMP procedure (SAS Institute) with restricted maximum-likelihood estimation. Restricted maximum-likelihood estimation produces estimates that are always positive and are generally more accurate than those produced by ANOVA methods.[19] Facets representing patients, examiners, study sites, and examiner experience levels were considered random effects. The facets representing examiner sex and examiner professional training were considered fixed effects. Because fixed effects are not generalizable, these 2 main effect facets were not considered further; however, the facets created by the interaction of fixed and random facets were random and were considered. All facets in the design were crossed, with the exception of examiners. Examiners were nested within a 4-way interaction (study site × examiner sex × experience level × professional group). From these variance components, an overall generalizability coefficient ($E\rho^2$) was calculated.

Results

Study 1: Internal Consistency

The means and standard deviations (SDs) for the various component scores obtained for this sample are shown in Table 1. The low means and standard deviations reflect the low prevalence of TMD in

Table 1 Internal Consistency of the CMI (n = 913)

CMI item group	Mean score	Standard deviation	Cronbach's alpha	No. of items
CMI total score	0.065	0.086	0.885	62
DI total score	0.059	0.064	0.517	26
Jaw mobility (+/–)	0.060	0.083	0.542	16
Jaw mobility (mm)			0.722	5
Joint sounds	0.105	0.158	0.109	4
Pain on TMJ palpation	0.026	0.096	0.636	6
MI total score	0.068	0.124	0.913	36
All muscles of mastication	0.061	0.123	0.882	24
Intraoral muscles of mastication	0.067	0.167	0.766	6
Extraoral muscles of mastication	0.060	0.126	0.855	18
All extraoral muscles	0.069	0.126	0.899	30
Neck muscles	0.082	0.161	0.832	12

Means and standard deviation are not reported for jaw mobility, because this would involve averaging more than 5 different jaw movement maneuvers.

this non-clinical sample. The internal consistencies of various CMI aggregate scores are also displayed in Table 1. Cronbach's alpha reached 0.885 for the CMI total score. The MI total score as well as summary scores for the extraoral masticatory muscles and the neck muscles showed reliabilities ranging from 0.832 to 0.913. Reliability for the intraoral muscles was slightly lower. In contrast, the internal consistency of the DI total score and its subcomponents was considerably lower, with the reliability of joint sound measurements being very poor. Jaw mobility measured on a dichotomous scale yielded an alpha value of 0.542, but this rose to 0.722 when measurements in millimeter units were used.

Correlations of Individual Items with Aggregate Scores. Correlations between the individual CMI items and the CMI, DI, and MI aggregate scores are shown in Table 2. For the CMI total score, most of the muscle and TMJ palpation items showed correlations with CMI total score in the 0.3 to 0.5 range. However, items measuring mandibular movement showed substantially lower correlations with the CMI total score, ranging from zero for jerky opening or closing movement to a maximum of 0.285 for pain on opening. Items intended to capture information about TMJ noises also showed low correlations with the CMI total score. For the CMI subscores, all muscle palpation items showed relatively strong correlations with the MI aggregate score. For the DI, jaw mobility items showed higher correlations with the DI score than they did with the CMI total score but remained in the 0.031 to 0.463 range. Items assessing TMJ palpation pain and TMJ noises showed relatively weak correlations with the DI total score.

Study 2: Interexaminer Agreement

The mean scores of female examinees were significantly higher than the corresponding scores of male subjects (all *P* values < .05). Mean scores (± SD) of women were: DI = 0.212 ± 0.171, MI = 0.328 ± 0.273, and CMI = 0.273 ± 0.208. Corresponding scores of men were: DI = 0.069 ± 0.040, MI = 0.045 ± 0.045, and CMI = 0.058 ± 0.041. Also, mean scores of Facial Pain Clinic patients were significantly greater than corresponding scores of healthy volunteers (all *P* values < .05). Mean scores of patients were: DI = 0.217 ± 0.159, MI = 0.324 ± 0.259, and CMI = 0.273 ± 0.202. Corresponding scores of healthy volunteers were: DI = 0.125 ± 0.125, MI = 0.152 ± 0.227, and CMI = 0.140 ± 0.174. There were no statistically significant differences in the scores assigned to patients by male versus female examiners or by dentist examiners versus non-dentist examiners. Furthermore, scores did not differ significantly as a function of the examiner's home site or experience level (backup, primary, or "gold standard"). These results argue against systematic bias introduced by these examiner characteristics.

Agreement on Aggregate Scores. The ICCs reflecting the reliability of CMI aggregate scores and some of their components are displayed in Table 3. The coefficients for CMI total score (ICC = 0.875), MI score (ICC = 0.856), and DI score (ICC = 0.808) all demonstrated good reliability.[20] The aggregate scores for joint sounds and TMJ palpation pain were somewhat lower.

Agreement on Individual Scale Items. The reliability of individual CMI items was also assessed

Table 2 Correlations Between Individual Items and Total Aggregate Scores for CMI, DI, and MI Scales (n = 913)

	Item total score correlations		
	CMI total	DI	MI
Mandibular movement			
Maximum opening	0.186	0.460	
Passive stretch opening	0.187	0.463	
Restriction on opening	0.125	0.386	
Pain on opening	0.285	0.396	
Jerky opening or closing	0.000	0.031	
S-deviation on opening or closing	0.063	0.195	
Lateral deviation on opening	0.099	0.182	
Pain on protrusive movement	0.223	0.325	
Limitation of protrusive movement	0.183	0.389	
Pain on right laterotrusion	0.237	0.276	
Limitation of right laterotrusive movement	0.255	0.449	
Pain on left laterotrusion	0.210	0.267	
Limitation of left laterotrusive movement	0.218	0.394	
Clinically can lock open	0.059	0.092	
Clinically can or is locked closed	0.134	0.244	
Rigidity of jaw on manipulation	0.134	0.306	
TMJ palpation			
Right lateral capsule	0.309	0.314	
Left lateral capsule	0.410	0.327	
Right posterior capsule	0.143	0.221	
Left posterior capsule	0.369	0.289	
Right superior capsule	0.326	0.316	
Left superior capsule	0.377	0.236	
TMJ noises			
Right click	0.082	0.186	
Left click	0.105	0.250	
Right crepitus	0.150	0.250	
Left crepitus	0.089	0.185	
Jaw muscles: Extraoral palpation			
Right anterior temporalis	0.405		0.416
Left anterior temporalis	0.455		0.476
Right middle temporalis	0.341		0.365
Left middle temporalis	0.424		0.468
Right posterior temporalis	0.417		0.425
Left posterior temporalis	0.388		0.416
Right deep masseter	0.541		0.553
Left deep masseter	0.518		0.535
Right anterior masseter	0.594		0.607
Left anterior masseter	0.547		0.565
Right inferior masseter	0.522		0.525
Left inferior masseter	0.609		0.620
Right posterior digastric	0.516		0.531
Left posterior digastric	0.535		0.531
Right medial pterygoid	0.470		0.494
Left medial pterygoid	0.558		0.589
Right vertex	0.316		0.338
Left vertex	0.364		0.391
Jaw muscles: Intraoral palpation			
Right lateral pterygoid	0.481		0.508
Left lateral pterygoid	0.500		0.527
Right medial pterygoid	0.486		0.508
Left medial pterygoid	0.484		0.513
Right temporalis insertion	0.521		0.529
Left temporalis insertion	0.489		0.530

Continued on page 290

Table 2 *(Continued)* Correlations Between Individual Items and Total Aggregate Scores for CMI, DI, and MI Scales (n = 913)

	Item total score correlations		
	CMI total	DI	MI
Neck muscle palpation			
Right superior sternocleidomastoid	0.343		0.365
Left superior sternocleidomastoid	0.372		0.411
Right middle sternocleidomastoid	0.544		0.572
Left middle sternocleidomastoid	0.564		0.589
Right inferior sternocleidomastoid	0.559		0.574
Left inferior sternocleidomastoid	0.516		0.548
Right trapezius insertion	0.512		0.538
Left trapezius insertion	0.464		0.511
Right upper trapezius	0.438		0.465
Left upper trapezius	0.490		0.531
Right splenius capitis	0.456		0.478
Left splenius capitis	0.576		0.600

Table 3 Reliability of CMI Scores (n = 106)

CMI score	ICC
CMI total score	0.875
MI score	0.856
Jaw muscles: Extraoral palpation	0.802
Jaw muscles: Intraoral palpation	0.671
Neck muscles	0.813
DI score	0.808
Jaw mobility	0.773
Joint sounds	0.634
TMJ palpation	0.666

ICCs shown here reflect the reliability of a single CMI examiner.

with a generalized kappa statistic, which is a chance-corrected percent agreement measure.[14] These results are presented in Table 4. The reliability of all items measuring muscle and joint palpation pain was approximately within the 0.4 to 0.6 range. The performance of examiners in evaluating mandibular movement depended greatly on which items were considered. Agreement was especially poor in evaluating jerky movement, S-deviation, lateral deviation on opening, jaw locking open or closed, and jaw rigidity on manipulation. Individual items measuring joint noises also demonstrated moderate reliability, with the notable exception of crepitus in the right joint, which was detected with a reliability of only 0.181. We are unable to offer any explanation as to why left and right crepitus were detected with such different reliabilities.

The CMI items relating to mandibular range of motion were measured on a ratio scale with a millimeter ruler and then converted to a binary categorical scale (positive or negative) prior to being tallied in the aggregate scores. Because there is a potential loss of information in converting from a ratio to a categorical scale, we calculated the ICCs separately for those items measured with a millimeter ruler. These results are displayed in Table 5. The reliabilities of the ratio scale measurements are consistently and appreciably higher than the corresponding categorical scale measurements (compare Table 5 with Table 4). A similar difference can be seen in Table 1 for internal consistency. Reliability did appear to suffer when measurements recorded in millimeters were converted to a categorical scale.

Agreement on Patterns of Signs and Symptoms. A difficulty with aggregate scores is that different examiners can arrive at the same aggregate score without agreeing on any individual items. Because all CMI items receive unit weights, all that is

Table 4 Multiple-Rater Kappas for Individual CMI Items (n = 106)

CMI item	Kappa
Mandibular movement	
Maximum opening	0.568
Passive stretch opening	0.551
Restriction on opening	0.528
Pain on opening	0.653
Jerky opening or closing	0.199
S-deviation on opening or closing	0.143
Lateral deviation on opening	0.240
Pain on protrusive movement	0.484
Limitation of protrusive movement	0.504
Pain on right laterotrusion	0.627
Limitation of right laterotrusive movement	0.340
Pain on left laterotrusion	0.585
Limitation of left laterotrusive movement	0.294
Clinically can lock open	0.008
Clinically can or is locked closed	0.192
Rigidity of jaw on manipulation	0.223
Jaw muscles: Extraoral palpation	
Right anterior temporalis	0.492
Left anterior temporalis	0.542
Right middle temporalis	0.467
Left middle temporalis	0.428
Right posterior temporalis	0.488
Left posterior temporalis	0.476
Right deep masseter	0.508
Left deep masseter	0.514
Right anterior masseter	0.520
Left anterior masseter	0.436
Right inferior masseter	0.480
Left inferior masseter	0.479
Right posterior digastric	0.433
Left posterior digastric	0.435
Right medial pterygoid	0.511
Left medial pterygoid	0.501
Right vertex	0.490
Left vertex	0.582
Jaw muscles: Intraoral palpation	
Right lateral pterygoid	0.402
Left lateral pterygoid	0.393
Right medial pterygoid	0.428
Left medial pterygoid	0.434
Right temporalis insertion	0.419
Left temporalis insertion	0.423
Neck muscle palpation	
Right superior sternocleidomastoid	0.434
Left superior sternocleidomastoid	0.510
Right middle sternocleidomastoid	0.577
Left middle sternocleidomastoid	0.620
Right inferior sternocleidomastoid	0.490
Left inferior sternocleidomastoid	0.491
Right trapezius insertion	0.450
Left trapezius insertion	0.451
Right upper trapezius	0.466
Left upper trapezius	0.449
Right splenius capitis	0.497
Left splenius capitis	0.450

Continued on page 292

Table 4 *(Continued)* Multiple Kappas for Individual CMI Items (n = 106)

CMI Item	Kappa
TMJ palpation	
Right lateral capsule	0.470
Left lateral capsule	0.491
Right posterior capsule	0.387
Left posterior capsule	0.414
Right superior capsule	0.525
Left superior capsule	0.505
TMJ noises	
Right click	0.494
Left click	0.462
Right crepitus	0.181
Left crepitus	0.548

Table 5 Reliability of Jaw Mobility Measured with a Millimeter Ruler (n = 106)

Measurement	ICC
Maximum voluntary opening	0.887
Maximum opening with passive stretch	0.883
Maximum protrusion	0.666
Maximum right laterotrusion	0.577
Maximum left laterotrusion	0.575

needed is that the examiners agree on the total number of items that should be scored positive. We assessed this possibility by employing R, a multivariate generalization of Cohen's kappa statistic, which quantifies the chance-corrected joint pattern of agreement across CMI items.[15,16] Because the sample of examiners varied from year to year, resulting in an unbalanced design, R was calculated for each year of the study, and these yearly values were averaged. The mean joint interexaminer agreement was $R = 0.278$ for all 62 CMI items, $R = 0.264$ for the 26 items contributing to the DI subscore, and $R = 0.319$ for the 36 items contributing to the MI subscore. Thus, when it was required that examiners agree on the specific pattern of signs and symptoms exhibited by a patient and not merely the total number of positive items, the reliability suffered dramatically (compare with ICCs for aggregate scores in Table 3).

Agreement with a "Gold Standard" Examiner. Another way to conceptualize examiner reliability is as the ability of a group of practicing examiners to identify and agree with a pattern of signs and symptoms identified by an expert examiner, who is designated a "gold standard." A multivariate, chance-corrected measure (R) was used to measure the agreement between the set of ratings produced by the study examiners and those produced by the "gold standard" examiner.[17] For all 62 CMI items, $R = 0.249$. For the 26 DI items, $R = 0.274$, and for the 36 MI items, $R = 0.318$. Thus, the study examiners generally were unable to replicate the global pattern of signs and symptoms identified by the "gold standard" examiner.

Study 3: Generalizability

The estimated variance components and their standard errors are displayed in Table 6. The first variance component (facet a in Table 6) represents variability resulting from differences among the subjects examined. This variance component is large relative to its standard error and relative to the other variance components. Next are shown the variance components representing specific sources of measurement error. These are small and collectively represent less than 10% of the total variance in DI, MI, and CMI scores. The final variance component (error a × f: [b × c × d × e] in Table 6) represents undifferentiated residual sources of measurement error that were not specifically identified by the present study design. The undifferentiated error component is larger than the components representing specific sources of error but is still considerably smaller than the component representing patients. Undifferentiated error explains 15.0% of DI variance, 8.5% of MI variance, and 7.5% of CMI variance. Individual differences among the patients examined account for the largest proportions of the variability (79.7% for DI, 83.9% for MI, and 85.8% for CMI), leaving only a small proportion to the other sources of measurement error investigated.

The generalizability coefficients ($E\rho^2$) for a single randomly selected examiner and/or the mean

Table 6 Variance Components and Standard Errors (n = 106)

	DI		MI		CMI	
Facet	Variance component	Standard error	Variance component	Standard error	Variance component	Standard error
a	0.024194	0.003549	0.063490	0.000084	0.039669	0.000032
b	0	0	0	0	0	0
c (fixed)	—	—	—	—	—	—
d (fixed)	—	—	—	—	—	—
e	0	0	0	0	0.000018	0.000194
f: (b × c × d × e)	0.000059	0.000126	0.000975	0.000703	0.000359	0.000260
a × b	0.000375	0.000824	0.002261	0.001035	0.001158	0.000587
a × c	0	0	0.000144	0.000671	0	0
a × d	0	0	0	0	0	0
a × e	0.000561	0.000496	0.001257	0.000902	0.001093	0.000539
b × c	0.000085	0.000150	0.000968	0.001215	0.000469	0.000498
b × d	0	0	0	0	0	0
b × e	0	0	0	0	0	0
c × d (fixed)	—	—	—	—	—	—
c × e	0	0	0.000090	0.000633	0	0
d × e	0	0	0	0	0	0
a × b × c	0.000498	0.001055	0	0	0	0
a × b × d	0	0	0	0	0	0
a × b × e	0	0	0.000076	0.001566	0	0
a × c × d	0	0	0	0	0	0
a × c × e	0	0	0	0	0	0
a × d × e	0	0	0	0	0	0
b × c × d	0	0	0	0	0	0
b × c × e	0	0	0	0	0	0
b × d × e	0	0	0	0	0	0
c × d × e	0	0	0	0	0	0
a × b × c × d	0	0	0	0	0	0
a × b × c × e	0	0	0	0	0	0
a × b × d × e	0	0	0	0	0	0
a × c × d × e	0	0	0	0	0	0
b × c × d × e	0	0	0	0	0	0
a × b × c × d × e	0	0	0	0	0	0
error a × f: (b × c × d × e)	0.004551	0.000712	0.006424	0.001289	0.003466	0.000472

Variance components less than 10^{-6} are listed as zeros.

a = Subjects examined (random); b = study sites (random); c = examiner sex (fixed); d = examiner professional training (fixed); e = examiner experience level (random); f = examiners (random).

taken over 2 examiners are shown in Table 7. These generalizability coefficients were interpreted as the ratio of universe score variance to expected observed score variance. They represented the generalizability of CMI scores used in making comparative decisions relative to other patients or some measure of group performance as opposed to decisions relative to a fixed criterion score. The coefficients were large, demonstrating that generalizability to a larger universe over the factors studied was high. Table 7 also contains estimates of generalizability coefficients for the case in which 2 examiners examined a patient and their scores were averaged. The generalizability of 2 examiners was numerically greater than that of a single examiner, but only marginally so.

Table 7 Generalization Coefficients (Ep^2) for a Single Examiner and for 2 Examiners Whose Scores Are Averaged (n = 106)

Scale	One examiner	Two examiners
DI	0.802	0.867
MI	0.862	0.901
CMI	0.874	0.909

Discussion

Previous reviewers have commented on the inherent variability of TMD signs and symptoms.[5] Symptoms may indeed change rapidly due to the natural history of the disorder, changes in jaw function, or even repeated palpation and mobility testing. In the present study, some patients began complaining after being subjected to repeated examinations. We randomized the order in which examiners examined patients to minimize systematic bias due to such sensitization effects, but this obviously can be an important source of examiner disagreement when multiple examinations occur over a short interval. On the other hand, lengthening the time between examinations also increases the likelihood that signs and symptoms will change due to the variable natural history of TMD. Previous studies suggest that joint sounds tend to vary even over short time intervals.[5] In fact, one of the signs of TMD measured on the CMI is a "nonreproducible" joint sound. Symptoms that do not occur reliably cannot be measured reliably.

Another issue that has not been resolved is what level of reliability is needed for dependable measurements. Previous facial pain researchers have recommended various acceptability criteria. Dworkin et al considered an ICC between 0.75 and 0.80 to be "acceptable."[5] Goulet et al, on the other hand, described ICCs in the 0.40 to 0.59 range as "moderate" and those in the 0.60 to 0.79 range as "good."[8] Such benchmarks are heuristically useful, but they are somewhat subjective and do not always provide a good impression of the impact that measurement error has on decisions drawn from the data. Much depends upon the use to which the data will be put. Furthermore, most investigators do not specify in enough detail the analysis model used or the assumptions in force. Shrout and Fleiss demonstrated that the same data could yield ICCs ranging from 0.17 to 0.91, depending on the assumptions in force, eg, whether a 1-way or a 2-way design is assumed, whether examiner variability is considered fixed or random, and whether the results are intended to apply only to a single examiner or to multiple examiners.[21]

Dworkin et al clearly demonstrated the value of retraining.[4] Interexaminer agreement increased significantly immediately following retraining. In our study, examiners were retrained annually in sessions lasting about 5 hours. In 1 previous study, examiners underwent 40 hours of training and calibration, and reliability was demonstrated to be high.[5] In yet another study, examiners were recalibrated every week.[8] Our results (studies 2 and 3) may overestimate reliability to some extent because assessments occurred soon after the retraining. Also, due to the logistical problems of assembling the various teams of raters at San Antonio each year, we were not able to assess test-retest reliability at, say, a 2-week interval. More research is needed to determine how much training is needed and what types of training and retraining are most effective.

The CMI developers reported very high ICC reliability for aggregate scores.[12] Dworkin et al reported comparably high reliability when summary scores, similar to those used on the CMI examination, were used.[5] We also found interexaminer agreement to be high when aggregate scores were considered. However, the data from both samples suggest that the reliability of the CMI might be improved by elimination or revision of certain items. For instance, items meant to measure joint sounds showed weak internal consistency (Table 1) and contributed little to the DI and CMI aggregate score variance (Table 2). Similarly, some of the items meant to measure the quality of jaw movement (S-deviation, lateral deviation, jerky movement, rigidity on manipulation, and locking open or closed) showed weak correlations with aggregate scores (Table 2), and examiners were not able to reliably detect their occurrence (Table 4). Future studies should address the impact of removing these items or substituting alternative items. If they are demonstrated to have high clinical importance, then we may need to concentrate our efforts on developing new measurement technologies. Our findings suggest that CMI reliability could be maintained or even possibly improved by eliminating some items. In the case of jaw mobility, there is a substantial loss of reliability in converting measurements made on a ratio scale to a binary (restricted versus unrestricted) categorical scale.

Reliability suffers when examiner agreement is defined as a pattern recognition task. The CMI examination, in conjunction with the training and calibration methods used in this study, did not yield highly reliable results when the standard was raised to this level. The practicing examiners were unable to agree among themselves on a specific pattern or constellation of signs and symptoms exhibited by patients. The practicing examiners also were unable to identify with high reliability the constellation of signs and symptoms identified as "correct" by the "gold standard" examiner.

Given these results, the high levels of reliability found for aggregate scores should not be greeted with great enthusiasm. Are the aggregate scores merely indexing the general severity of the disorder? Are a few items with very low reliability limiting examiners' ability to agree on symptom profiles? Further research is needed to answer these questions and to see whether training methods can be developed that will improve pattern recognition accuracy.

The application of generalizability theory did not identify specific sources of measurement error that are likely to be problematic in future studies. In well-trained examiners, gender, professional training, experience level, and geographic distribution do not play a major role in limiting the dependability of aggregate CMI scores. The largest variance component was that representing undifferentiated error not accounted for by the present study design. Future generalizability studies may identify key sources of measurement error that can be directly addressed during examiner training. On the other hand, the undifferentiated error term may represent the collective effects of a multitude of minor forces. In this case, only comprehensive examiner training will be successful.

Reliability can be raised by increasing the number of examiners and averaging their scores. This is the solution usually recommended by generalizability theory. However, given the potential for sensitization of patients following repeated examinations, it may not be a workable solution. Our results (see Table 7) suggest that the gain in reliability expected by increasing the number of examiners from 1 to 2 and averaging their scores is modest and probably not worthwhile.

While few would disagree that examiner reliability should be maximized, this study and others reviewed above suggest that the standard of consistently high reliability has not always been met. Even when reliability can be shown to meet "acceptable" standards, most data analysts ignore the fact that interexaminer agreement is imperfect. A preferable analysis strategy might be to include a factor representing examiner uniqueness in the statistical analysis model so that the extent of examiner agreement can, in some measure, be taken into account.[22]

Acknowledgments

This work was supported by research grants P50DE10756 and R01DE09630.

References

1. Fricton JR, Schiffman EL. The craniomandibular index: Validity. J Prosthet Dent 1987;58:222–228.
2. Helkimo M. Studies on function and dysfunction of the masticatory system. Index for anamnestic and clinical dysfunction and occlusal state. Swed Dent J 1974;67:101–121.
3. Dworkin SF, LeResche L. Research diagnostic criteria for temporomandibular disorders: Review, criteria, examinations and specifications, critique. J Craniomandib Disord Facial Oral Pain 1992;6:301–355.
4. Dworkin SF, LeResche L, DeRouen T. Reliability of clinical measurement in temporomandibular disorders. Clin J Pain 1988;4:89–99.
5. Dworkin SF, LeResche L, DeRouen T, Von Korff M. Assessing clinical signs of temporomandibular disorders: Reliability of clinical examiners. J Prosthet Dent 1990;63:574–579.
6. Eriksson L, Westesson P-L, Sjoberg H. Observer performance in describing temporomandibular joint sounds. J Craniomandib Pract 1987;5:33–35.
7. Goulet J-P, Clark GT. Clinical TMJ examination methods. Calif Dent Assoc J 1990;18:25–33.
8. Goulet J-P, Clark GT, Flack VF, Liu C. The reproducibility of muscle and joint tenderness detection methods and maximum mandibular movement measurement for the temporomandibular system. J Orofacial Pain 1998;12:17–26.
9. Brennan RL. Elements of Generalizability Theory. Iowa City: American College Testing Program, 1983.
10. Crocker L, Algina J. Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston, 1986.
11. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. The dependability of behavioral measurements: Theory of generalizability of scores and profiles. New York: Wiley, 1972.
12. Fricton JR, Schiffman EL. Reliability of a craniomandibular index. J Dent Res 1986;65:1359–1364.
13. Dahlstrom L, Keeling SD, Fricton JR, Galloway Hilsenbeck S, Clark GM, Rugh JD. Evaluation of a training program intended to calibrate examiners of temporomandibular disorders. Acta Odontol Scand 1994;52:250–254.
14. Fleiss JL. Statistical methods for rates and proportions. New York: John Wiley and Sons, 1981.
15. Berry KJ, Mielke PW Jr. A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. Educ Psychol Meas 1988;48:921–933.
16. Berry KJ, Mielke PW Jr. A generalized agreement measure. Educ Psychol Meas 1990;50:123–125.
17. Berry KJ, Mielke PW Jr. Measuring the joint agreement between multiple raters and a standard. Educ Psychol Meas 1997;57:527–530.
18. Bravo G, Potvin L. Estimating the reliability of continuous measures with Chronbach's alpha or the intraclass correlation coefficient: Toward the integration of two traditions. J Clin Epidemiol 1991;44:381–390.
19. Marcoulides GA. An alternative method for estimating variance components in generalizability theory. Psychol Reports 1990;66:379–386.
20. Landis JR, Kock GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
21. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 1979;86:420–428.
22. Gilthorpe MS, Maddick IH, Petrie A. Introduction to multilevel modeling in dental research. Community Dent Health 2000;17:222–226.